Functional transcription factor target discovery via compendia of binding and expression profiles

SCIENTIFIC REPORTS

OPEN

Functional transcription factor target discovery via compendia of binding and expression profiles

Received: 19 June 2015
Accepted: 07 January 2016
Published: 09 February 2016

Christopher J. Banks[1], Anagha Joshi[2] & Tom Michoel[1]

Genome-wide experiments to map the DNA-binding locations of transcription-associated factors (TFs) have shown that the number of genes bound by a TF far exceeds the number of possible direct target genes. Distinguishing functional from non-functional binding is therefore a major challenge in the study of transcriptional regulation. We hypothesized that functional targets can be discovered by correlating binding and expression profiles across multiple experimental conditions. To test this hypothesis, we obtained ChIP-seq and RNA-seq data from matching cell types from the human ENCODE resource, considered promoter-proximal and distal cumulative regulatory models to map binding sites to genes, and used a combination of linear and non-linear measures to correlate binding and expression data. We found that a high degree of correlation between a gene's TF-binding and expression profiles was significantly more predictive of the gene being differentially expressed upon knockdown of that TF, compared to using binding sites in the cell type of interest only. Remarkably, TF targets predicted from correlation across a compendium of cell types were also predictive of functional targets in other cell types. Finally, correlation across a time course of ChIP-seq and RNA-seq experiments was also predictive of functional TF targets in that tissue.

Transcriptional regulation by DNA-binding transcription-associated factors and chromatin modifiers (here collectively abbreviated as "TFs") is a fundamental process determining all aspects of cell behaviour, and TFs are known to be essential for a wide range of important cellular and organismal phenotypes. Using ChIP-sequencing technology[1], the ENCODE and modENCODE consortia have generated detailed maps of the genomic locations where TFs bind in diverse human cell types[2] and in important model organisms[3–5]. Invariably, these experiments have demonstrated that TFs bind throughout the genome over a quantitative continuum of occupancy levels[6] and that the number of binding events can significantly exceed the number of known or possible direct target genes[7]. Hence, predicting when the binding of a TF in a gene locus will lead to a biologically significant change in the rate of transcription of the neighbouring gene ("functional DNA binding", see the glossary of terms by Biggin[6]) is a challenging and largely unsolved problem.

Several studies have recently used ChIP-sequencing data of TFs and/or histone modifications to predict absolute expression levels in a particular cell type[8–11]. While these studies show that a large proportion of the variation in expression levels across genes can be explained by the presence or absence of TF-binding sites for particular combinations of TFs[8,9], this approach is ill-suited to predict functional TF targets, i.e. to predict *differential* gene expression in a particular cell type upon perturbation of the TF. Indeed, a recent large-scale study where 59 TFs where knocked down in a human lymphoblastoid cell line (GM12878) concluded that only a small subset of genes bound by a factor within 10 kb of their transcription start site (TSS) were differentially expressed following knockdown of that factor[12]. However, Cheng *et al.*[9] also showed that differential TF binding between two cell types correlates with differential gene expression between those two cell types, suggesting that functional TF target genes can possibly be predicted through the "guilt-by-association" principle by correlating TF-occupancy and gene expression levels across multiple cell types.

In other applications of genomics, function is often predicted in this manner. Genes with similar expression pr[14], genetic interaction profiles[15] or protein interaction partners[16] often share the same molecular function. Likewise, putative DNA-regulatory motifs are identified from their shared occurrence in the upstream regulatory sequences of co-expressed genes[17], networks of TF-regulatory interactions are inferred by associating TF activity

[1]Division of Genetics and Genomics, The Roslin Institute, The University of Edinburgh, UK. [2]Division of Developmental Biology, The Roslin Institute, The University of Edinburgh, UK. Correspondence and requests for materials should be addressed to A.J. (email: Anagha.Joshi@roslin.ed.ac.uk) or T.M. (email: Tom.Michoel@roslin.ed.ac.uk)

profiles to candidate target expression levels[18–20], and long-range DNA contact interactions between regulatory elements and putative target genes can be predicted by correlating open chromatin (measured by sequencing DNase I hypersensitive sites[21]) and gene expression levels across multiple cell types[22–26].

To test the hypothesis that the guilt-by-association principle applies to the discovery of functional TF target genes, we used ChIP- and RNA-sequencing data across multiple cell types from the human ENCODE resource[2]. We considered several cumulative regulatory models to map ChIP-peaks to genes, ranging from proximal binding to incorporating distal events: 1 kb/5 kb/10 kb/50 kb around the TSS, the nearest TSS, and 1 kb/5 kb around the TSS and in the gene body. This is consistent with the emerging view that TF-binding sites act redundantly to promote robustness against genetic and environmental perturbations, and that they may regulate their target genes in a cumulative manner[27]. Furthermore, since there exists no gold standard data of functional DNA-binding events (in the sense defined above), we used the knockdown data of Cusanovich *et al.*[12] as a proxy measurement. Five of the knocked down factors had ChIP-seq binding maps available in at least ten cell types in the human ENCODE data, with matching RNA-seq gene expression data, and were considered in this study. Using three different correlation measures, individually and in combination, we found that the correlation between variation in cumulative binding around the TSS of a gene and variation in expression levels of that gene was a better predictor of functional effects than the presence of multiple binding events in the cell type where the TF knockdown was performed. Remarkably, these results were confirmed when using correlation across a time course of ChIP-seq and RNA-seq experiments during mouse circadian rhythm[28].

Methods

Preparation of data. The binding events (peaks) for eight transcription-associated factors (CEBPB, EP300, EZH2, MYC, RAD21, REST, TAF1 and YY1) with binding profiles in ten or more cell types were downloaded from the ENCODE resource[2]. Peaks were mapped to transcription start sites (GENCODE v12) using seven different models (1 kb/5 kb/10 kb/50 kb around the TSS, the nearest TSS, and 1 kb/5 kb around the TSS and in the gene body). To calculate the peak height at each peak, we downloaded the corresponding mapped read files (BED files) for each sample from the ENCODE resource. We then calculated the coverage using BEDTOOLs and normalized coverage count was used as an estimate for peak height. The RPKM values of RNA sequencing data (GENCODE v12) for the corresponding cell lines were also downloaded from the ENCODE resource. They were then quantile normalised using R to be comparable across samples. ENCODE binding and expression profiles were available for 24,392 genes. The global expression change (in the form of differential expression *P*-value s) upon deletion of five of the eight factors (EP300, EZH2, RAD21, TAF1 and YY1) in a lymphoblastoid cell line (GM12878) was available for 8,872 genes (also called the "reference genes" below)[12]. Differentially expressed genes upon deletion of CEBPB were obtained directly from the Gene Expression Omnibus using the GEO2R tool (accession number: GSE54975). Differentially expressed genes upon deletion of MYC were obtained from Seitz *et al.*[29].

The binding events (peaks) for six circadian regulators (Bmal1, Clock, Cry1, Cry2, Per1, Per2) and two RNA polymerase II states (8WG16 and ser5p) at six time points (1 hr, 4 hr, 8 hr, 12 hr, 16 hr and 20 hr) as well as the RPKM values of RNA sequencing data at the same time points from murine liver samples were obtained from Koike *et al.*[28]. A total of 2629 genes had both binding and RPKM data available for Per2. Differentially expressed genes upon deletion of Per2 in murine liver[30] were obtained directly from the Gene Expression Omnibus using the GEO2R tool (accession number GSE30139; 2409 genes with fold change >1.5 and FDR corrected *P*-value <0.05).

From each source of data, we obtained a number of working sets of data where for each TF there exists concordant data from binding, expression, and knockdown. For each TF and each peak-to-gene model there were three datasets. The ChIP and expression data were matrices with a row for each gene and a column for each condition, containing respectively the number of peaks mapped to the TSS of that gene and its expression level. The knockdown data for ENCODE was an expression change *P*-value for each gene and for mouse circadian was a list of genes with significant expression change.

We also prepared ChIP data for the same factors with quantitative binding information (sum of peak magnitudes for each gene). Finally we prepared alternative datasets filtered for CpG-rich and CpG-depleted promoters, as obtained from the UCSC genome browser.

Prediction of functional TF target genes. The approach we took to predicting functional target genes looked at the correlation between binding and expression profiles over a range of cell-types or conditions. However, we found that correlation by any one method alone is a not necessarily a good predictor for any given factor or binding model. We found that different correlation methods identify different types of relation between binding and expression. To improve prediction we combined results from a number of correlation methods in a wisdom of crowds approach.

For each dataset (i.e. for each TF and each peak-to-gene model) we computed the correlation between the number of binding peaks and the expression of each gene across all conditions, using three correlation measures: the absolute Pearson correlation coefficient (PC), the absolute Spearman correlation coefficient (SC), and the absolute combined angle ratio statistic (CARS). CARS is our variant of the angle ratio statistic (ARS) of Marstrand and Storey[25]. The ARS was shown to have high power for detecting associations when both variables are restricted to a narrow relative range, with one or very few cell types appearing as distinct outliers[25]. Whilst ARS only considers positive associations between variables (i.e. where an increased number of binding sites corresponds to increased expression of a target gene), CARS uses the same principle to test for both positive and negative associations. We define CARS as follows. Vector $\vec{x}$ is a vector of RNA-seq data and $\vec{y}$ is a vector of ChIP-seq data (for the same cell-types). Both vectors have length t. Both vectors are then scaled: $\vec{x}^s = \frac{\vec{x}}{\max(|\vec{x}|)}$ and likewise for $\vec{y}^s$. Both

vectors are then median centred: $\vec{x}^* = \vec{x}^s - \text{med}(\vec{x}^s)$ and likewise for $\vec{y}^*$. Outlier distance is measured for each point:

$$\vec{d} = \left(\sqrt{x_1^{*2} + y_1^{*2}}, \ldots, \sqrt{x_t^{*2} + y_t^{*2}}\right) \tag{1}$$

and the ratio statistic

$$\vec{r} = \left(\frac{d_1}{\text{med}(\vec{d})}, \ldots, \frac{d_t}{\text{med}(\vec{d})}\right) \tag{2}$$

quantifies the distance of each point from the medoid. We then take the angle of each point from the x-axis $\vec{\theta} = (\theta_1, \ldots, \theta_t)$ and form a positive angle statistic $\vec{\Delta}^+ = (\Delta_1^+, \ldots, \Delta_t^+)$, the angular distance of each point from the line $x = y$, and a negative angle statistic $\vec{\Delta}^- = (\Delta_1^-, \ldots, \Delta_t^-)$, the angular distance of each point from the line $x = -y$:

$$\Delta_i^+ = \begin{cases} |45 - \theta_i| & 0 \le \theta_i < 135 \\ |225 - \theta_i| & 135 \le \theta_i < 315 \\ |45 - (\theta_i - 360)| & 315 \le \theta_i < 360 \end{cases}, \quad \Delta_i^- = 90 - \Delta_i^+ = \begin{cases} |315 - (\theta_i + 360)| & 0 \le \theta_i < 45 \\ |135 - \theta_i| & 45 \le \theta_i < 225 \\ |315 - \theta_i| & 225 \le \theta_i < 360 \end{cases}. \tag{3}$$

This is where CARS differs from ARS, which only measures angular distance from the line $x = y$. The positive scores for each point are $ARS^+ = (ARS_1^+, \ldots, ARS_t^+)$ where $ARS_i^+ = r_i \times e^{c\Delta_i^+}$ and the negative scores ARS^- are formed in the corresponding manner, with $c<0$ a fixed parameter of the method. The overall combined angle ratio statistic is then defined as

$$CARS = \begin{cases} ARS_{\max}^+ & ARS_{\max}^+ \ge ARS_{\max}^- \\ -ARS_{\max}^- & ARS_{\max}^+ < ARS_{\max}^- \end{cases} \tag{4}$$

where $ARS_{\max}^{\pm} = \max(ARS^{\pm})$. The value of the parameter c was determined by requiring that empirical P-values satisfied a correct null distribution (i.e. such that P-values >0.5 had a uniform distribution), following the procedure of Marstrand and Storey[25]. In this study we used $c = -0.01$ which conformed to this requirement for all data sets.

For all three correlation measures, we kept track of whether the score originated from a positive or negative association. For comparison purposes, we also considered the number of binding peaks in the cell type where the knockdown experiment was performed (called "multiple binding") as a functional target predictor.

Measures of performance. We used the knockdown data as our gold standard for defining a functional effect of TF-binding on target gene expression. We measured the predictive performance of each correlation method and multiple binding by computing the precision vs. recall (PR) curves (where *precision* is the proportion of predicted genes that are in the gold standard and *recall* is the proportion of the gold standard set that was predicted) for each dataset. We then took, for each dataset, thresholds on both the level of multiple binding and the correlation scores, and genes with scores above the threshold were taken as positive predictions. To achieve comparable results between different TFs, we chose thresholds to give a specific fold increase in precision over the background proportion of bound genes with a knockdown effect (i.e., the proportion of genes bound in the knockdown cell type that have a significant functional effect in the knockdown). We recorded the positively identified genes and calculated the hypergeometric overlap P-value between the predicted gene set and gold standard set. We chose to make comparisons at a 1.5-fold precision increase over the background. In other words, if a fraction f of the reference gene set were differentially expressed upon knockdown of a particular TF, we determined thresholds for the PC, SC and CARS such that a fraction $1.5f$ of the genes exceeding the threshold were differentially expressed. For making predictions for TFs for which no gold standard set was available, we chose the threshold to be the top 1% of predictions.

Performance was measured on the intersection of genes with available knockdown differential expression data (the reference gene set) and real correlation scores (i.e., non-constant binding and expression profiles) or multiple-binding score (i.e., at least one peak in the knockdown cell type for the current peak-to-gene model); in the CARS method, outliers in one dimension (e.g. with constant binding profile) were penalized by the angular penalty but not excluded from the calculation, as recommended by Marstrand and Storey[25] to ensure a correct null distribution. For the biological validation of the predicted target sets, all genes in the complete set of 24,392 genes above the given score threshold were used, irrespective of the availability of knockdown data.

Results

Number of binding sites in a gene locus is a weak predictor of functional relevance of binding. In order to test the hypothesis that correlation between TF binding and gene expression across cell types can be used as a predictor of the functional relevance of binding events, we used data from the human ENCODE resource[2]. We selected eight TFs, each with ChIP-sequencing profiles in at least ten cell lines and corresponding RNA-sequencing data available for the same cell lines. We first assigned genome-wide binding locations (peaks) to putative target genes if they were within 5 kb of the transcription start sites (TSSs) obtained from GENCODE V12. We then defined the binding profile of a gene as the number of peaks associated to that gene across the

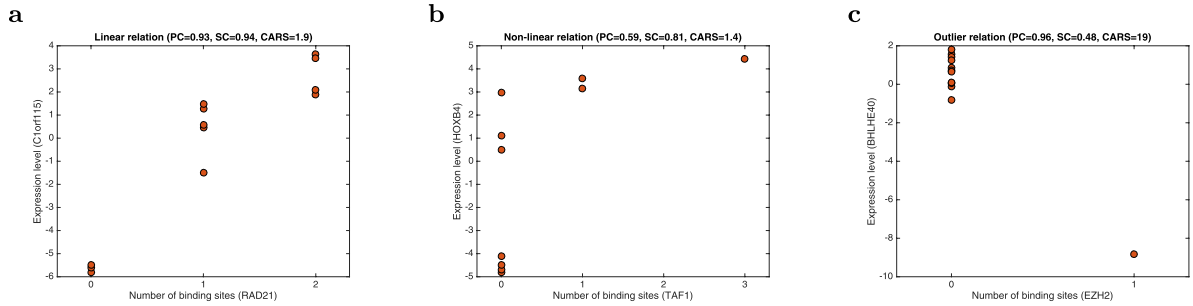

Figure 1. Characteristic scatter plots of binding and expression profiles for known differentially expressed targets, showing a linear relation favoured by Pearson correlation (**a**), a non-linear monotonic relation favoured by Spearman correlation (**b**), and an outlier relation favoured by CARS (**c**).

available cell types. We defined functional relevance of binding of a TF in a gene locus by whether or not the gene is differentially expressed upon knockdown of the transcription factor. Although a limited and stringent definition of functional relevance, this facilitated us to make a systematic and quantitative comparison of different functional prediction methods. For five of the selected TFs (EP300, EZH2, RAD21, TAF1 and YY1), siRNA knockdown data in a lymphoblastoid cell line (GM12878) was available in the form of differential expression *P*-values for 8,872 genes[12]. The differentially expressed genes (*P*-value <0.05) were considered functional targets of the corresponding transcription factor (true positives) and the non-differentially expressed genes true negatives. This resulted in five datasets (one for each TF) with binding profiles for 24,392 genes, together with matching expression profiles for the same genes in the same cell types as well as true positive and true negative functional target gene lists (see Methods for details).

We first calculated for each TF the ratio of functional targets among genes bound by the TF in GM12878. These ratios showed only a very limited increase compared to the genome-wide background ratio of functional targets, which was significant for only two factors (EP300 and RAD21, hypergeometric *P*-value <0.05, Supplementary Table 1). This result is consistent with the finding by Cusanovich *et al.*[12] that only 12 of their 29 knockdowns with available ChIP-sequencing data resulted in a significant overlap between binding and differential expression. We then tried to improve results by predicting targets only if multiple binding sites are present. For three factors (EP300, RAD21, and YY1), we could achieve a 1.5-fold increase in precision (i.e., percentage of known functional targets) with hypergeometric *P*-value <0.05. However, the threshold number of binding sites had to be large and consequently the predicted target sets were small (Supplementary Table 1).

Correlation across compendia of binding and expression profiles improves functional target prediction. To investigate the utility of correlation-based methods to predict functional targets, we considered three distinct association measures between binding and expression profiles across multiple cell types: the absolute Pearson correlation coefficient (PC), which tests for a positive or negative linear association (Fig. 1a), the absolute Spearman correlation coefficient (SC), which tests for a positive or negative monotonic trend (Fig. 1b), and the combined angle ratio test statistic (CARS), which tests for a non-linear "on-off" relationship in a positive or negative direction (Fig. 1c) (see Methods).

We validated the predictive performance of each method for each TF independently using precision-recall (PR) curves. In all cases, high-scoring genes for any of the correlation measures were more likely to be differentially expressed upon knockdown of the TF than low-scoring genes (Fig. 2). In particular, for four out of five factors (EP300, EZH2, RAD21, YY1) a small number of targets were predicted with a precision close to one. This enrichment (high-precision-low-recall region) translated into an improvement in the number of significant sets and the size of set predicted compared to using multiple binding in the knockdown cell type. To enable comparison of different methods across different factors, we determined for each method and each factor the score threshold resulting in a 1.5-fold increase in precision compared to the genome-wide background ratio of functional targets, and calculated the significance of the overlap between all functional targets and targets predicted at this threshold using a hypergeometric test. We found significantly enriched target sets for each factor for at least one, and often multiple, of the methods, predicting significantly more target genes compared to using binding data from the knockdown cell type only (MB) (Fig. 3 and Supplementary Table 2). Interestingly, taking the union of the predicted gene sets for each method gave a further increase in overlap with the gold standard (Fig. 3 and Supplementary Table 2). This was explained by the fact that enriched target sets predicted by each method (PC, SC, and CARS) showed only a limited overlap (Fig. 4a), suggesting that all types of relations (linear, non-linear monotonic, on-off; cf. Fig. 1) do occur between binding and expression profiles of functional TF targets.

Since all three methods take into account both positive and negative associations between binding and expression, we asked whether specific TFs show a bias for either sign. For four out of five factors (EP300, RAD21, TAF1 and YY1), positive interactions dominated the predicted target sets, suggesting that they mostly function as activators of expression. In contrast, for EZH2 most predicted interactions were negative (Fig. 4b), consistent with the fact that more than two thirds of the differentially expressed genes were up-regulated in response to EZH2 knockdown[12]. EZH2 is indeed known to repress transcription by participating in histone mark H3K27me3 deposition as well as DNA methylation[31], although it also functions as a double-faceted molecule in breast cancers, either as a transcriptional activator or repressor of NF-kB targets, depending on the cellular context[32].

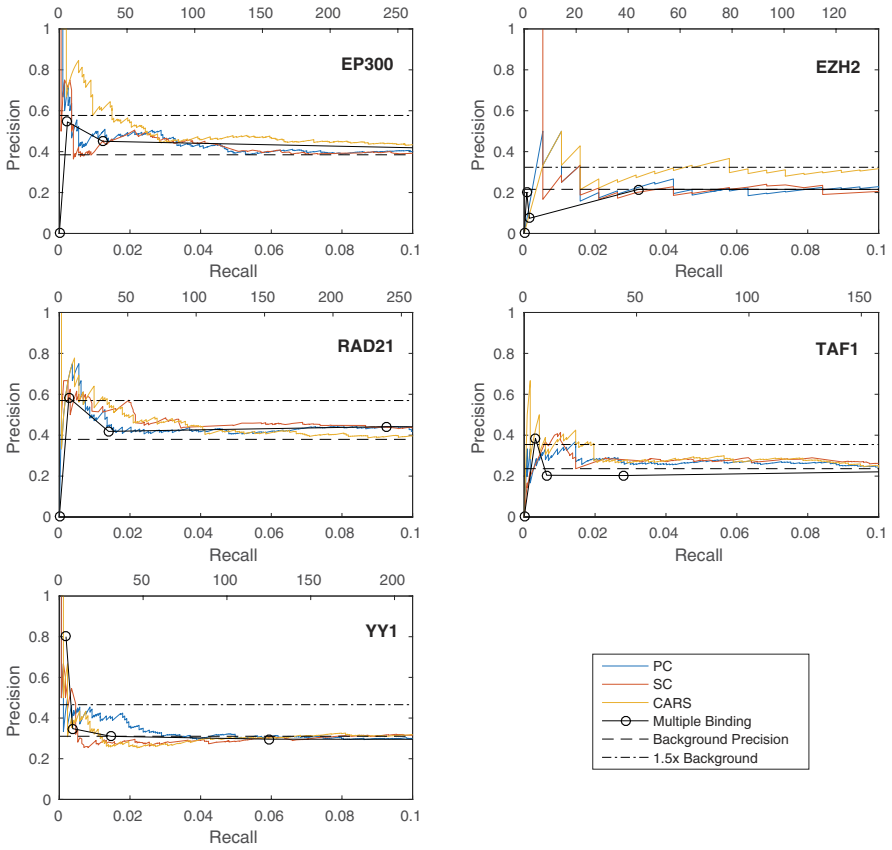

Figure 2. Precision vs. Recall curves for functional binding of EP300, EZH2, RAD21, TAF1 and YY1 in the 5 kbTSS peak to gene model predicted by each method, using 8,872 reference genes (true positive and true negative functional targets). The top scale on the x-axis shows the number of true positives for the corresponding recall value.

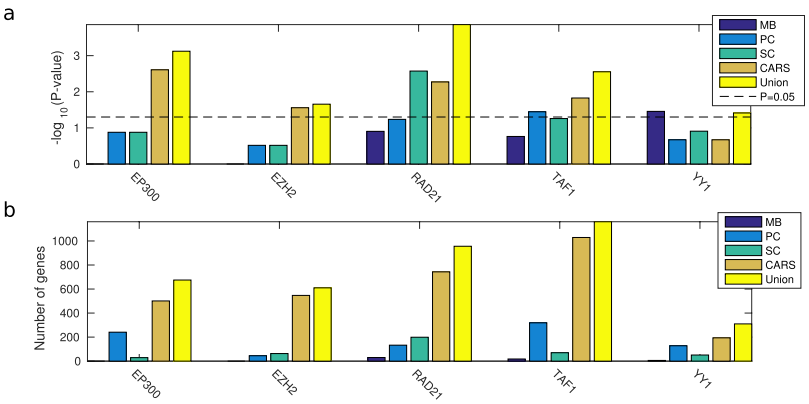

Figure 3. Functional target set enrichment significance (negative log hypergeometric *P*-value) (**a**) and size (**b**) for each method and TF at a predicted 1.5-fold precision over background. This is compared to using binding data from the knockdown cell type only (MB). Significance was determined by analysing the 8,872 gold standard reference genes, whereas sets sizes refer to subsets of the complete set of 24,392 genes with correlation score exceeding the 1.5-fold precision threshold derived from the gold standard.

Cell-type specific studies have previously used quantitative TF binding information (i.e. peak heights) to predict absolute expression levels[8,9]. We therefore also tested our method using ChIP-seq peak heights in place of

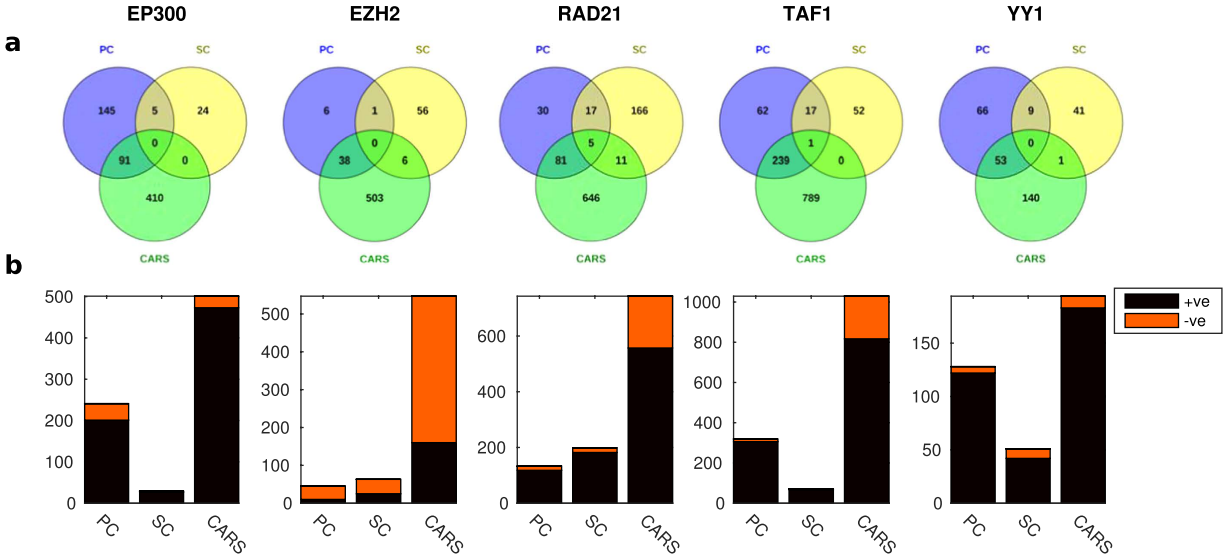

Figure 4. Overlap of functional target sets for each factor predicted by each method at a predicted 1.5-fold precision over background (**a**). Number of positively and negatively correlated targets predicted by each method for each factor at a predicted 1.5-fold precision over background (**b**). Set sizes in both panels refer to subsets of the complete set of 24,392 genes exceeding the 1.5-fold precision threshold determined by analysing the 8,872 reference genes.

counts. However we found no significant improvement in prediction using quantitative information. Precision levels varied, but were similar to results obtained via on-off binding (see Supplementary Figures 3 and 4—compare with Fig. 3). Cheng *et al.*[9] also found that the cell-type specific expression of CpG-rich promoters was easier to predict than CpG-depleted promoters. We therefore partitioned the gene set in the same way, and compared results. Here too, the results showed that there is no significant difference on the precision overall between CpG-rich and CpG-depleted genes (see Supplementary Figure 5), even after correcting for the difference in the number of CpG-rich and CpG-depleted genes (see Supplementary Figure 6). However some improvement is shown for EZH2; this is not unreasonable, given that EZH2 is known to target CpG-island promoters[33].

To understand the functional relevance of the predicted target sets, we performed gene ontology enrichment analysis using BiNGO[34]. The target gene sets of each factor (Supplementary Data 1) were enriched for distinct and specific functional categories, such as "inflammatory response" (EP300; P-value $< 1.2 \times 10^{-7}$), "generation of neurons" (EZH2; P-value $< 1.6 \times 10^{-12}$). In contrast, the sets of genes bound a factor in the knockdown cell type (Supplementary Data 2) were enriched for metabolic processes for all factors. Furthermore, of the 2,618 predicted functional target genes across all five TFs, 70% were predicted to be target of only one factor. In contrast, 82% of the 5,764 bound genes were bound by more than one TF. Taken together, these results suggest that the correlation-based method is able to select smaller and more specific sets of functional target genes from the hundreds to thousands of genes bound by a given factor in a given cell type.

Full target sets are available in Supplementary Data sets 59–93.

Correlation between binding and expression predicts functional targets in non-ENCODE cell types. In the previous analyses, the gold standard validation data (differential expression results) were obtained in a cell type that was also present as one of the ENCODE cell lines used to predict targets. Next we asked whether correlation of binding and expression across a compendium of cell types is also informative for predicting targets in cell types not present in the compendium.

Firstly, we predicted functional targets for three additional transcription factors (MYC, CEBPB and REST) where binding and expression information was available in ten or more ENCODE cell types, but not the perturbation data. These TFs typically bind to a few thousand gene loci in any given cell type, but the correlation-based method enabled us to select only a few hundred high-confidence functional targets by taking the union of top 1% predictions from each method. The predicted targets of CCAAT/enhancer binding protein beta (CEBPB) were specifically enriched for the Wnt signalling pathway (P-value $= 7.9 \times 10^{-4}$). CEBPB has a demonstrated role in the suppression of Wnt/β-catenin signaling during adipogenesis[35]. CEBPB targets were also enriched for the functional category "positive regulation of cytokine production during immune response" (P-value $= 9.1 \times 10^{-2}$), in line with the well characterised role of CEBPB in the regulation of immune and inflammatory response genes[36]. Since the ENCODE cell types are cancer-related rather than immunological or adipogenesis-related, this demonstrates the strength of the correlation-based approach to find the most functionally relevant targets of a transcription factor. Similarly, the predicted targets of REST were enriched for functions specific to neurons, such as presynaptic membrane (P-value $= 4.1 \times 10^{-3}$). REST is identified as a key regulator to protect neurons in parts of the brain from oxidative stress, as well as protein aggregations characteristic of many neurodegenerative diseases[37].

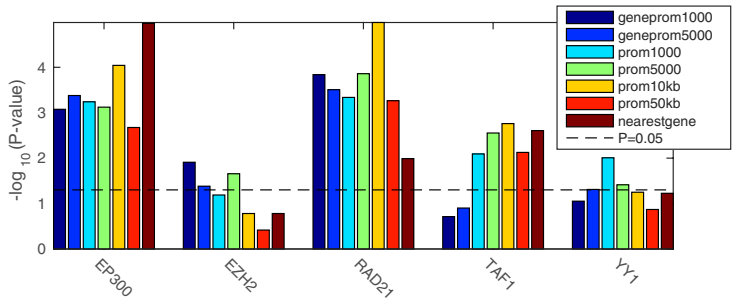

Figure 5. Functional target set significance (hypergeometric P-value) predicted by the union of all correlation methods for all peak-to-gene models at a predicted 1.5-fold precision over background. Set sizes refer to subsets of the complete set of 24,392 genes exceeding the 1.5-fold precision threshold determined by analysing the 8,872 reference genes.

Next, having confirmed that the correlation-based method is able to identify the most functionally relevant targets, we investigated whether the high-confidence predicted targets are also valid in other cell types. Reschen *et al.*[38] systematically investigated the role of CEBPB in macrophage differentiation in an *in vitro* model for coronary artery disease. By performing ChIP sequencing for CEBPB before and after macrophages differentiate into foam cells, they identified 5866 genes where CEBPB was bound at significantly higher levels in foam cells. Of these differentially bound genes, 16% (935) were differentially expressed between foam cells and macrophages. Of the 749 predicted CEBPB targets using our correlation-based method across ten ENCODE cell lines, also 16% (119) of genes were differentially expressed between foam cells and macrophages. Similarly, Seitz *et al.*[29] studied the role of MYC in Burkitt Lymphoma (BL) and identified 7054 MYC binding sites (6169 within 5 kb of a TSS) in 5 BL cell lines. 530 (8.5%) of these bound genes were differentially expressed after siRNA-mediated knock-downs of MYC in BL cell lines. Our method predicted 728 MYC targets using the ENCODE data, which showed 9% overlap with genes differentially expressed after siRNA-mediated knock-downs of MYC in BL cell lines. These two analyses demonstrate that predictions derived using our approach have a precision that is comparable to the ChIP sequencing performed in the exact cell line of interest. The recall on the other hand is limited at this point due to the limited availability of cell types with matching binding and expression data to build functional target predictions, and the possibility of missing cell type specific targets.

Full target sets are available in Supplementary Datasets 3–58.

Different binding target models work better for different TFs. To investigate whether assigning genome-wide binding locations (peaks) to putative target genes using different peak-to-gene models results in improved performance, we considered seven peak-to-gene models: 1 kb, 5 kb, 10 kb or 50 kb around the transcription start sites (TSSs) obtained from GENCODE V12; 1 kb or 5 kb around the TSS and within the gene body; or the nearest TSS. We then predicted the targets using the union of correlation methods for each of the gene models. We noted that different peak-to-gene association models performed best for different factors (Fig. 5). Promoter proximal binding has previously been associated with the functional relevance of binding[39]. Accordingly, for all five TFs, the 5 kbTSS correlation gave significant enrichment of functional targets (Fig. 5). EZH2 and YY1 performed best when TSS proximal peaks were considered. This suggests that EZH2 and YY1 are involved in transcriptional control through promoter proximal elements only. To note, both EZH2 and YY1 predominantly bind in promoter proximal regions[2]. Given that overall the 5 kbTSS model performed better than the 1 kbTSS model, this suggests the presence of alternate cell type specific promoters within 5 kb of the consensus annotated promoter. For EP300, the farther the peaks from the TSS were considered, the better the predictions, such that the best results were obtained when all peaks were associated to the nearest TSS (Fig. 5). As the presence of EP300 is associated with active enhancers[40], it is not surprising to find that the nearest TSS model works best for EP300. Similarly, the 10 kb peak-to-gene model was the best performer for RAD21.

Among the three correlation measures, CARS predicted the highest number of targets at the 1.5-fold precision threshold for most peak-to-gene models, with the exception of the nearest gene model for EP300, where PC performed significantly better, indicating that linear interactions dominate in this case (Supplementary Figures 1 and 2).

Correlation across time points of a defined biological process is also predictive of functional effects. We have demonstrated that correlation between binding patterns and expression values of genes across multiple cell types can be used to predict functional targets. We then investigated whether the correlation-based approach can be extended to other data types, such as time series. The mammalian circadian clock is a cell-autonomous process with a period of about 24 hours. It controls the sleep-wake cycle, blood pressure and hormone secretion by regulating key processes such as metabolism, the cell cycle and DNA repair through feedback loops of transcriptional regulators (Clock, Bmal1, Cry1, Cry2, Per1 and Per2) essential for the rhythmicity[41]. We obtained genome-wide binding patterns (ChIP-sequencing data) for the six transcription factors listed above, as well as gene expression profiles (RNA-sequencing data) at six time points (1, 4, 8, 12, 16 and 20 hours) in murine liver[28]. We predicted high confidence correlated genes by considering the top 1% predictions of the union of correlations between binding and expression across this time series, using the 5 kbTSS promoter

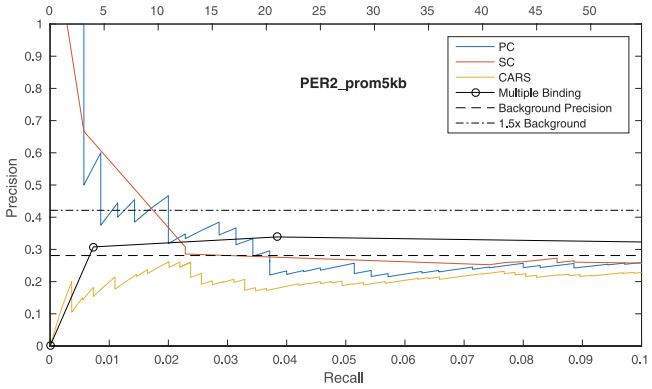

Figure 6. Precision vs. Recall curves for functional binding of PER2 in the 5 kbTSS peak-to-gene model by each method, using 21200 reference genes (true positive and true negative functional targets).

proximal peak-to-gene model. A gold standard set of genes differentially expressed upon TF-perturbation in liver was available only for Per2[30]. As before, the PR curve showed that targets predicted by correlation, but not multiple binding, were enriched for known functional targets (Fig. 6).

The high-confidence predicted Per2 targets were highly expressed at 16 and 20 hours, similar to the gene expression pattern of Per2 itself, and were more likely to be bound by Cry1/Cry2 than Clock/Bmal1[28]. They were enriched for the functional category "regulation of RNA metabolic process" (P-value $= 4.3 \times 10^{-3}$). Although 10% of the liver transcriptome follows a circadian rhythm, only about half of it can be explained by de novo transcription, suggesting that mRNA processing may play a key role in the circadian rhythmicity. Per2 is associated with RNA processing through the RNA-dependent methylation process[42]. This again demonstrates that the correlation-based approach enables the identification of a small set of highly-relevant functional targets among the tens of thousands of genes bound by a given transcription factor.

Discussion

In this study, we have applied the guilt-by-association principle to predict functional targets of transcription-associated factors by testing if a gene's TF-specific binding profile across multiple cell types correlated (positively or negatively) with its expression profile across the same cell types, using three distinct correlation measures (Pearson and Spearman correlation and the combined angle ratio statistic) and a range of cumulative regulatory models for mapping TF-binding peaks to transcription start sites. Compared to the traditional approach where target genes are inferred from the presence of one or more binding sites in a gene locus in a cell type of interest, the three correlation-based methods showed improved prediction of functional targets, defined here as genes differentially expressed upon TF knockdown, especially when used in combination.

It is known that TFs function in a condition-specific manner, and it may not be obvious that correlation-based measures across multiple cell types are able to identify functional targets. However, it is precisely the presence of binding and associated change in a target gene's expression level in the cell type(s) where the TF is active, and the absence of this signal in other cell types, which leads to a high-confidence prediction. The angle ratio statistic was developed precisely to detect such cell-type specific effects with high specificity[25], and was indeed found to predict a significantly higher number of functional targets at the same enrichment threshold compared to the Pearson and Spearman correlation, which predominantly select linear or monotonic trends, respectively. Furthermore, if binding of a factor varies across a compendium of cell types, then the correlation between binding and expression was found to be predictive of functional effects even in cell types that were not part of the compendium. Although more work is needed to investigate the condition-specific properties of predicted functional target genes, our results suggest that correlation-based predictions capture both condition-specific and condition-independent targets of a TF.

Interestingly, these results were confirmed using a time course of matching binding and expression data in a single cell type, showing the wide validity of the guilt-by-association principle for functional TF-binding prediction. In contrast to the ENCODE results, only the Pearson and Spearman correlation predicted significantly enriched target sets in this case, whereas the CARS outlier detection method did not perform well. This is consistent with the fact that samples from the same tissue at different time points are more similar to each other than samples from highly distinct cell types, and emphasizes the importance of combining different correlation methods to detect all types of signal present in a dataset.

Four limitations affect the current study and need to be taken into account. Firstly, our definition of a gold standard of true positive and true negative functional target genes from differential expression data following knockdown of a TF is only a proxy for true functional binding events, namely when the binding of a TF in a gene locus significantly affects the gene's rate of transcription. However no large-scale data of changes in transcription rates following TF knockdowns is currently available. Secondly, although the human ENCODE ChIP-seq matrix currently reports data for nearly 200 TFs and more than 80 cell types, it is very sparse. Indeed, only eight TFs had ChIP-seq profiles available in more than ten cell types with matching RNA-seq data, which we considered a minimum to perform a correlation-based analysis. Of the eight factors considered, half were sequence-specific TFs (CEBPB, MYC, REST and YY1) and half were general factors: two promoter-associated (EZH2 and TAF1),

one enhancer-associated (EP300) and one involved in three-dimensional DNA organization (RAD21). Of the sequence-specific TFs, only one (YY1) had knock-out data available in the lymphoblastoid cell line and could thus be validated directly. As more data will become available, it will be important to establish if the reported results also hold for a wider range of sequence-specific transcription factors. Thirdly, predicting the effect of a particular TF on a particular gene naturally depends on the reliability of the ChIP-seq experiments for that TF, but even within the ENCODE resource, with its high standards for technical quality control, the biological quality of samples is not always guaranteed[43]. Lastly, we only considered the presence of binding sites and expression data to investigate the improvement in functional prediction by correlation-based methods compared to using the presence of binding sites only. Although we found that taking into account binding peak height or promoter CpG content did not improve our predictions, we did not consider the presence of sequence motifs or various chromatin features. These have been shown to improve prediction of cell-type specific variation in expression among genes[8,9], and will likely also improve prediction of functional targets. Having established the validity of the correlation-based method, future work will be aimed at building functional target gene predictors that combine this approach with additional data types and existing knowledge. Despite these limitations, we believe that the use of correlated features in compendia of binding and expression profiles with matching conditions is a powerful novel method to predict functional TF target genes, which is able to identify high-confidence functional target genes among the thousands of genes bound by a given transcription factor.

References

1. Park, P. J. ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics* **10,** 669–680 (2009).
2. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
3. Gerstein, M. *et al.* Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science* **330,** 1775–1787 (2010).
4. Roy, S. *et al.* Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* **330,** 1787–1797 (2010).
5. Yue, F. *et al.* A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515,** 355–364 (2014).
6. Biggin, M. D. Animal transcription networks as highly connected, quantitative continua. *Developmental Cell* **21,** 611–626 (2011).
7. MacQuarrie, K. L., Fong, A. P., Morse, R. H. & Tapscott, S. J. Genome-wide transcription factor binding: beyond direct target regulation. *Trends in Genetics* **27,** 141–148 (2011).
8. Ouyang, Z., Zhou, Q. & Wong, W. H. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *PNAS* **106,** 21521–21526 (2009).
9. Cheng, C. *et al.* Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Research* **22,** 1658–1667 (2012).
10. Dong, X. *et al.* Modeling gene expression using chromatin features in various cellular contexts. *Genome Biology* **13,** R53 (2012).
11. Budden, D. M. *et al.* Predicting expression: the complementary power of histone modification and transcription factor binding data. *Epigenetics & Chromatin* **7,** 1–12 (2014).
12. Cusanovich, D. A., Pavlovic, B., Pritchard, J. K. & Gilad, Y. The functional consequences of variation in transcription factor binding. *PLoS Genetics* **10,** e1004226 (2014).
13. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *PNAS* **95,** 14863–14868 (1998).
14. Hughes, T. R. *et al.* Functional discovery via a compendium of expression profiles. *Cell* **102,** 109–126 (2000).
15. Costanzo, M., Baryshnikova, A., Myers, C. L., Andrews, B. & Boone, C. Charting the genetic interaction map of a cell. *Current Opinion in Biotechnology* **22,** 66–74 (2011).
16. Sharan, R., Ulitsky, I. & Shamir, R. Network-based prediction of protein function. *Molecular Systems Biology* **3,** 88 (2007).
17. Roth, F. P., Hughes, J. D., Estep, P. W. & Church, G. M. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mrna quantitation. *Nature Biotechnology* **16,** 939–945 (1998).
18. Friedman, N. Inferring cellular networks using probabilistic graphical models. *Science* **308,** 799–805 (2004).
19. Bussemaker, H. J., Foat, B. C. & Ward, L. D. Predictive modeling of genome-wide mRNA expression: from modules to molecules. *Annu Rev Biophys Biomol Struct* **36,** 329–347 (2007).
20. Balwierz, P. J. *et al.* ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. *Genome research* **24,** 869–884 (2014).
21. Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132,** 311–322 (2008).
22. Xi, H. *et al.* Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome. *PLoS Genetics* **3,** e136 (2007).
23. Natarajan, A., Yardmc, G., Sheffield, N., Crawford, G. & Ohler, U. Predicting cell-type-specific gene expression from regions of open chromatin. *Genome Research* **22,** 1711–1722 (2012).
24. Sheffield, N. C. *et al.* Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome Research* **23,** 777–788 (2013).
25. Marstrand, T. T. & Storey, J. D. Identifying and mapping cell-type-specific chromatin programming of gene expression. *PNAS* **111,** E645–E654 (2014).
26. Demeyer, S. & Michoel, T. Graph-based data integration predicts long-range regulatory interactions across the human genome. *arXiv preprint arXiv:1404.7281* (2014).
27. Spivakov, M. Spurious transcription factor binding: Non-functional or genetically redundant? *BioEssays* **36,** 798–806 (2014).
28. Koike, N. *et al.* Transcriptional architecture and chromatin landscape of the core circadian clock in mammals. *Science* **338,** 349 (2012).
29. Seitz, V. *et al.* Deep sequencing of myc dna-binding sites in burkitt lymphoma. *PLoS ONE* **6** (2011).
30. Zani, F. *et al.* Per2 promotes glucose storage to liver glycogen during feeding and acute fasting by inducing *Gys2 PTG* and G_L expression. *Molecular Metabolism* **2,** 292–305 (2013).
31. Viré, E. *et al.* The Polycomb group protein EZH2 directly controls dna methylation. *Nature* **439,** 871–874 (2006).
32. Lee, S. T. *et al.* Context-specific regulation of NF-κB target gene expression by EZH2 in breast cancers. *Molecular Cell* **43,** 798 - 810 (2011).
33. Riising, E. M. *et al.* Gene silencing triggers polycomb repressive complex 2 recruitment to cpg islands genome wide. *Molecular cell* **55,** 347–360 (2014).
34. Maere, S., Heymans, K. & Kuiper, M. Bingo: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21,** 3448–3449 (2005).
35. Chung, S. S. *et al.* Regulation of wnt/β-catenin signaling by CCAAT/Enhancer Binding Protein β during adipogenesis. *Obesity* **20,** 482–487 (2012).
36. Pruitt, K. D. *et al.* RefSeq: an update on mammalian reference sequences. *Nucleic Acids Research* **42,** D756–D763 (2014).

37. Lu, T. *et al.* REST and stress resistance in ageing and alzheimer/'s disease. *Nature* **507,** 448–454 (2014).
38. Reschen, M. E. *et al.* Lipid-induced epigenomic changes in human macrophages identify a coronary artery disease-associated variant that regulates PPAP2B expression through altered C/EBP-Beta binding. *PLoS Genetics* **11,** e1005061 (2015).
39. Whitfield, T. *et al.* Functional analysis of transcription factor binding sites in human promoters. *Genome Biology* **13,** R50 (2012).
40. Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457,** 854–858 (2009).
41. Lowrey, P. L. & Takahashi, J. S. Genetics of circadian rhythms in mammalian model organisms. *Advances in Genetics* **74,** 175 (2011).
42. Fustin, J.-M. *et al.* RNA-methylation-dependent RNA processing controls the speed of the circadian clock. *Cell* **155,** 793–806 (2013).
43. Devailly, G., Mantsoki, A., Michoel, T. & Joshi, A. Variable reproducibility in genome-scale public data: A case study using ENCODE ChIP sequencing resource. *FEBS letters* doi: 10.1016/j.febslet.2015.11.027 (2015).

Acknowledgements

A.J. is a Chancellor's Fellow of the University of Edinburgh. This work was supported by Roslin Institute Strategic Grant funding from the BBSRC.

Author Contributions

C.B. performed the analysis and wrote the manuscript. A.J. and T.M. conceived the idea, collected the data, performed the analysis and wrote the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Banks, C. J. *et al.* Functional transcription factor target discovery via compendia of binding and expression profiles. *Sci. Rep.* **6**, 20649; doi: 10.1038/srep20649 (2016).